DOCUMENT RESUME

ED 461 259 FL 024 539

AUTHOR Spolsky, Bernard
TITLE Prognostication and Language Aptitude Testing, 1925-62.
PUB DATE 1994-09-00
NOTE 26p.; In: Language Aptitude Invitational Symposium Program Proceedings (Arlington, VA, September 25-27, 1994); see FL 024 538. Based on chapter from author's book: "Measured Words," published by Oxford University Press, March 1995. Edited version published in "Language Testing," 12(3), p321-340, 1995.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Admission Criteria; Educational History; Foreign Countries; *Individual Differences; *Language Aptitude; *Language Tests; Measurement Techniques; *Prediction; Second Language Programs; Student Placement; *Testing
IDENTIFIERS *Modern Language Aptitude Test

ABSTRACT

Objective public testing methods were brought into use in language programs in the 1930s, primarily to justify decisions to exclude unqualified students from high school foreign language classes. In the United States, after World War II, government language programs supported research into the assessment of language aptitude to improve selection techniques for expensive, intensive language training. While one such study by a group of psychologists failed, another led to development of usable language aptitude tests, the Modern Language Aptitude Test. This study concluded that language aptitude consisted of four distinct, measurable abilities: phonetic coding; grammar handling; rote memorization of a large number of vocabulary items; and inductive language learning ability. It also added three important dimensions to prognosis: measurement of some of the components of individual variation in language aptitude; a model showing how measurable abilities interact with goals and methods; and illustration that aptitude was only one of the factors involved in the general theory of language learning. (Contains 33 references.) (MSE)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 461 259

PROGNOSTICATION AND LANGUAGE APTITUDE TESTING -- 1925-62

Bernard Spolsky
Language Policy Research Center, Bar-Ilan University and the National Foreign Language Center, Johns Hopkins University

It is a signal honor to have been invited to address this Symposium, a further important milestone in the long-established collaboration between academic language testers and the government language teaching and testing establishment.[1] While the first interest in language aptitude came from the colleges and universities in the 1920s, the major developments in language aptitude testing in the 1960s were the result of government initiative, and it is most fitting that CALL should have taken the lead in this intended to continue the refinement of the field.

Language testing is a field that has long recognized its social and political significance. A hundred years before Foucault, in the brilliantly stimulating few pages he devoted to examinations, showed their disciplinary effect in providing 'un regard normalisateur, une surveillance qui permet de qualifier, de classer et de punir' (Foucault, 1975:186-7),[2] Henry Latham (1877) was already decrying the "encroaching power" of examinations that, he protested, was biasing education, blurring important distinctions between liberal and technical education, and narrowing the range of learning through forcing students to prepare by studying with crammers and in cramming schools.

Ironically, examinations had long been regarded as forces for good and a method of attaining to equal opportunity. The original Chinese system, that lasted two thousand years, was intended to recruit civil servants on the basis of their excellence rather than

---

[1]This paper was based on a chapter from my book, *Measured Words*, published by Oxford University Press in March 1995. Research on it was carried out while I was on sabbatical leave from Bar-Ilan University as a Mellon Fellow at the National Foreign Language Center. It has been revised to be the opening plenary paper at the 1994 Language Aptitude Invitational Symposium sponsored by the Center for the Advancement of Language Learning, held at Rosslyn, VA, from September 25-27, 1994. This is the version prepared for the meeting. An edited version has also been published in *Language Testing*, 12 (3) 321-340, 1995.

[2]'a normalizing gaze, a surveillance that makes it possible to qualify, to classify and to punish'.

FL024539

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Bernard Spolsky
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

their patronage, and it was this 'Chinese principle' that was used by Lord Macaulay to bolster arguments for using examinations for selecting cadets for admission to the India Civil Service that was one of the major reforms in nineteenth century England. The egalitarian potential of the public examination no doubt contributed to its importance in the United States after independence and in Revolutionary France, although clearly Napoleon saw its potential for centralized control.

It was concern with the fairness of powerful public examinations that led Edgeworth (1888) to call attention to their "unavoidable uncertainty." The new-type objective test was seen as a solution to this problem. Objective testing started to increase in Britain and the United States in the decade or so after the First World War, but only in America did it find an immediate public acceptance, as the testing business started to sweep American education in the late 1920s

Language testing was not immune to objectivisation. By 1930, the work of the Modern Language Study had demonstrated that the achievement test or examination could be a powerful tool for control over the language teaching process, and in the hands of the College Entrance Examination Board, the proficiency test or examination was developing into an equally effective way to maintain authority over the language qualifications of applicants for admission to universities or countries. Between the World Wars, these tests evolved steadily, with constant progress towards objectivization and industrialization that I have discussed elsewhere (Spolsky, 1995).

There remained another area of disquiet, the control of admission to the language learning class itself, and it is with this parallel development that this Symposium a[illegible] this paper that opens it will deal. In the first half of the paper, I will describe attempts made in the U.S. between the two world wars to develop prognosis tests with the goal of ensuring that only qualified students would be allowed into high school language classes. In the second, I will describe two research programs, one a failure and the other a major success, to develop aptitude tests that would allow government agencies to select only appropriate candidates for expensive intensive language training.

There are two main points that this study will reveal. The first is that the level of success of the efforts was more a function of the resources made available to the task than of the state of knowledge or sophistication of the researchers. The pre-World War II enterprise of language teachers to control access to their classes was local and conducted with minimal funds; nonetheless, useful tests were developed and a general theoretical model of considerable sophistication was established.[3] The later government and foundation supported undertaking, encouraged by the Cold War and government concern for the cost of intensive language instruction, led to two major studies, one of which reached a much higher level of practical usefulness.

The send point is that, by the late thirties, it was widely and clearly understood that aptitude, however defined and however precisely measured, could only account for part of the variance in language learning success. The fuller instructional model set out by Carroll (1962), but understood in general terms at least thirty years earlier, show clearly that the various kinds of aptitude interacted with other personal factors (such as motivation) and with the instructional conditions to produce various kinds of success in language learning. In fact, by the 1930s, all of the items that might be included in this fuller model had been mentioned, so that the task was not to think of new ones, but to show the contribution of each to the model.

Prognosis testing

Our story then starts some sixty or more years ago. While egalitarian principles demanded that everyone should have the right of access to a high school education, including foreign language classes that were offered in them, the tiny amount of time allocated in the US school curriculum to language study led to a distressingly high failure rate. Motivated by what Michel (1936) referred to as 'the deplorable mortality in foreign language classes,' language testers set out to develop what they called prognosis tests,

[3] In fact, most if not all the ideas proposed at the Symposium as relevant to aptitude had been mentioned before 1942; what had not been done had been to show the exact weight to be given to each feature, but the Symposium papers did not do this either.

which, they hoped, could provide information about how well someone would perform in a language learning situation, or more precisely, about how to keep prospective failures out of their classes.

The genesis of prognosis tests was strictly practical rather than theoretical. Once it had become accepted in the USA in the early 1920s that general intelligence tests could be used with some effect to forecast how well a student would do at school, it was inevitable that some people would start to ask about the possibility of predicting success in specific subjects, including language study. This could then be used to alleviate the problems of teachers who felt themselves required to deal with students they believed unqualified for language study and who had been admitted to their classes through a policy of mass education.

This concern was highlighted in a paper entitled 'Mortality in modern languages students' by Cheydleur (1932a) reporting a long-term study of drop-outs and failures in language classes at the University of Wisconsin. After painting a picture of language departments agonizing over the numbers of their students who dropped out or failed their courses, Cheydleur argued for the value of using intelligence, placement and advancement tests to control student access and progress. Between 1925 and 1930, three prognosis tests for school use were prepared that stayed on the market for many years.

From the beginning, these tests combine two separate approaches to testing aptitude, which might be labeled the **analytical** and the **synthetic.** The analytical approach was to use items that tapped specific hypothesized cognitive abilities, usually through the first language, such as memory or vocabulary or some other aspect of verbal intelligence.

The synthetic approach was to give the candidate a mini-lesson in an artificial or foreign language, assuming that one could generalize from this short experience to performance in longer learning programs.

One of the earliest tests was written by Stoddard and Vander Beke, which included six subtests, three involving English grammatical skills -- singulars and plurals, tense,

nominalization, and three to do with guessing Esperanto words, applying Esperanto grammatical rules, and translating Esperanto sentences into English. A second was the *Language Aptitude Test* prepared by a team at George Washington University (Hunt et al., 1929), which involved learning elements of an artificial language. A third was the *Luria-Orleans Modern Language Prognosis Test* (Luria and Orleans, 1928), which took 85 minutes contained a language learning trial, consisting of vocabulary exercises (cognates and memorization) and eight grammar translation lessons in French and Spanish.

Prognosis in the Modern Foreign Language Study

It was while these early tests were being developed that the field of foreign language teaching was subjected to a major review by the Modern Foreign Language Study and the Canadian Committee on Modern Languages that started work in 1924 and went on for some years. The members of the committee were ardent supporters of prognosis:

> This Committee felt that no part of its experimental program would be more welcome to its colleagues as likely to throw light on their problems and bring relief from the difficult and often hopeless situation created by the numbers and unfitness of students, and it arranged, therefore, as soon as the foreign language achievement tests were well under way, to sponsor experimental undertakings in the field of prognosis. (Henmon, 1929:v).

The motivation was fundamentally economic, the goal being to replace 'wasteful methods of trial and error' with more efficient selection of students and their assignment 'to the work for which they are best fitted.' (Henmon 1929:3) The problems studied by the eight researchers whose work was supported and reported by the Committees turned out in the event, however, to be extremely resistant to solution and their studies were discouragingly inconclusive. In the long run, they failed to

> bring evidence that any test has yet been devised which can be counted on to reveal linguistic incapacity or to show itself as a reliable instrument for selecting

successful students of foreign languages. The question of language prognosis is far too complex for such a categorical answer. (Henmon 1929:vi)

> The theoretical question underlying the design of a selection technique was whether the mind should be conceived of as a 'host of highly specialized capacities which may vary independently' or as 'a unitary affair' with the various parts correlated and forming 'a common factor of general intelligence.' American educational psychology, Henmon noted, was inclined to the belief in a high degree of specialization, which was why the search for specific abilities was being so enthusiastically pursued. He saw the task as being to determine the relative contributions of general intelligence and special aptitudes to predictions about student performance.

The belief in the importance of special aptitude was well entrenched in the profession. Two-thirds of the US and Canadian modern language teachers questioned in a 1926 survey had found cases of students with 'linguistic disability or incapacity not accompanied by low general intelligence.' Intelligence was believed to be a factor. Henmon saw it as the task of his research group to answer four basic questions:

- Is there a minimal IQ level for successful modern language study?
- Is there a minimal general scholarship level for successful modern language study?
- Can special language learning abilities be recognized, tested, and used for prediction of success?
- Can one semester's results be used to predict future success?

The work of the 1920s was reported in a book published by the Modern Language Study (Henmon *et al.* 1929). In the introductory essay, Henmon summarized recent work looking at correlations of intelligence quotients and scores with school marks or objective test scores in modern languages. Most of the studies had shown a low positive correlation,

ranging from 0.20 to 0.60, not much use for practical decision making. The 'variability, inaccuracy and subjectivity of school marks' were so well established that they could not be expected to help much. But Henmon was convinced of the value of the continued search for special language abilities.

In the first of six reports of current work, John Bohan looked at the relation between scores on intelligence tests given to entering students at the University of Minnesota between 1921 and 1925 and their later grades in English and Foreign Languages, finding correlations between 0.15 and 0.50.

Carl Brigham, teaching at Princeton University and already associated with the College Entrance Examination Board where he was developing the SAT, studied the Princeton artificial language test invented by Stuart Dodd, which had been shown to have high diagnostic validity as a general intelligence test but the prognostic adequacy of which was limited. Brigham analyzed various correlations in the case of 236 men for whom there were full enough data. The best predictor of college French marks was the average of College Board Entrance Examinations in French, English and Latin (0.480); neither the intelligence test (0.276) nor the language test (0.269) were nearly as useful predictors, nor did the latter two tests add much to the prediction of the examinations (0.533).

In another chapter, L. Thomas Hopkins, at the University of Colorado, found the Wilkins Prognosis Test and the Wilkins Elimination test to be 'a reliable measure of some kind of ability or particular type of function,' but not of the ability to succeed in foreign languages.

George Rice, at the University of California, gave a test written by May Barry which taught some Spanish grammar items and vocabulary to 100 pupils as a trial experience in language learning. The test correlated with intelligence quotient (0.79), and with teacher's marks at the end of the year (0.60) better than the intelligence score did (0.53).

Percival Symonds, teaching at Teachers College, Columbia University, whose test was later used in a number of studies and must have been widely accepted, pointed out the problem of determining the value of a prognostic test. Even if such a test could measure aptitude, it was judged by its correlation with achievement, which was the combined result of aptitude, 'and of the forces of instruction, including interest and interest of the learner, organization of the material, skill of the teacher, etc.' This model, set out formally by Carroll (1960, 1962) is often forgotten or overlooked by researchers venturing into the area of aptitude testing for the first time. None of the earlier researchers ever claimed that aptitude alone accounted for research; there is nothing novel in the claims (heard even at the 1994 Symposium) that other personal or instructional factors need to be taken into account.

But recognizing this complexity made the validation of a prognosis test doubly difficult: first, a test will have normally been used to exclude unsuitable students from the course and so from the validation study, and secondly, the aptitude test is known to measure and account for only part of the assumed causes of later variation.

In spite of this problem, Symonds believed that three types of aptitude tests made sense: measures of general intelligence, tests of ability in the student's native language, and 'quick-learning tests in the new language.' He gave pupils in four schools a set of intelligence tests compiled by E.L. Thorndike, four quick-learning tests (two by Dodd and two using Esperanto by Symonds himself) and the Iowa Placement Examination (Foreign Language Aptitude). Those pupils who lasted the semester then took the American Council Beta French and Spanish Tests. While various problems with the skewing of some of the tests meant that the regression weights could not be relied on, the correlations suggested that those tests which included elements of translation ability (grammatical knowledge in particular) were likely to be good predictors of success in the classes.

In the final chapter, John Todd, a psychologist at the University of California, included in a test items based on a psychological analysis of the language learning process: a general questionnaire, a test of immediate auditory memory span for isolated digits, and

tests of the extent of native vocabulary and range of information. A number of studies were carried out. IQ was found to correlate well with school marks in languages. IQ tests also correlated well with Todd's linguistic test. Todd was satisfied that he had not found evidence of a special language aptitude: 'Whatever our tests may have measured it plainly was not a linguistic "talent" or special aptitude. If linguistic special aptitude is a reality, some other distinct type of test must be invented for the purpose of measuring it.' (1929:161) Todd's negative findings must have had a temporarily dampening effect on what ultimately proved to be the most useful avenue of research, namely the testing of much more specific abilities.

With the publication of the collection of papers by Henmon *et al.* (1929), the place of prognosis as a central topic in language testing research had been established, and the general model within which a solution must be found had been delineated, but there had been no widely accepted answers to the questions that had been raised.

The Symonds tests of prognosis

Over the next decade, research on prognosis continued. Symonds continued his research with the aptitude test that he had designed (1930a), reporting in a study (1930b) a correlation of 0.71 between the prognosis test and a later achievement test.

The effectiveness of the Symonds' Foreign Language Prognosis Tests was examined in a number of studies over the next few years. Richardson (1933) administered them to 242 high school freshmen planning to take foreign languages finding a correlation of about 0.60 with first semester scores. Richardson did find the prognosis tests gave better predictions than intelligence tests with two cohorts of 120 high school students.

In research for an MA thesis at the University of Chicago, Lau (1933) administered the test to eighty pupils in three Michigan high schools on their first day of class, and found a 0.60 correlation with the American Council Alpha tests at the end of the semester. The weakest correlation was with vocabulary and the strongest with grammar.

An elaborate study using both the Symonds' and the Iowa foreign language aptitude tests was undertaken as a master's thesis at the University of Minnesota, Sister Virgil Michel (1934, 1936), a teacher at the St. Joseph's Academy. She acknowledged her inspiration to the statement by Symonds that 'prognostic testing is the romantic chapter in the history of educational measurement,' and agreed also with the platitude that failing students should have been guided into easier classes, but noted that educational prognosis was 'still in its infancy.'(1936:275) She administered the Symonds Foreign Language Aptitude Test to a group of high school students and the Iowa Foreign Language Test to a smaller group of beginning college German students at the college level) and to both, a newly devised German prognosis test that she had constructed including a memory test of short German sentences with their English translations, an analogies test of words that were cognate in German and English, and a series of German grammar rules and exercises. For the high school students, none of these tests gave useful correlations with the Columbia Research Bureau German Test or with teachers' marks at the end of the first semester. Multiple correlations combining the tests did not help much. She concluded that the Symonds test using Esperanto seemed to have not done as well with German as with French and Spanish For the university students, combinations of the Iowa test (which also used Esperanto) and the German prognosis tests did achieve correlations with the end of semester marks, but not much better than did the high school average. Her thesis concludes somewhat pessimistically:

> In general, the experiment corroborates the findings of the majority of investigators in foreign language prognosis in so far as the correlations are rather low, in so far as predicting success in any one subject is much more difficult than prognosis of success in all subjects in high school or university, and in so far as it points with increasing insistence to the need for further research to secure more efficient predictive measures than those that exist at present. (Cited from Coleman and King, 1938:435).

'Romantic' as the topic may have been, there were no signs of a happy ending yet, but the Symonds Test continued to produce useful results for French high school classes,

and Sister Virgil's recognition of the possible language specificity was an important advance

Kaulfers on prognosis

If the correlations cited so far seem low, an even more pessimistic picture emerged from the work of the California foreign language education researcher, Walter Kaulfers (1931), who found IQ scores or English marks to be better predictors than standardized foreign-language aptitude tests. Kaulfer's work on prognosis formed the basis of his Ph.D. dissertation written at Stanford University (1933). Reviewing over 650 correlations, published since 1901 by nearly fifty researchers, between foreign language achievement and nearly seventy other factors, he found large variability. The medians for the most common factors were prognosis tests (0.60), English ability (0.46), general language ability (0.44) and mental ability (0.35). His work left Kaulfers unconvinced that there was a special language aptitude, and he judged the prognosis tests to be nothing more than weighted intelligence tests. Because of the unstandardized conditions in junior high school Spanish classes, he saw little likelihood of getting predictive efficiency of much higher than twenty to thirty percent. As early as Kaulfer's dissertation, then, it was fully understood that the effectiveness of an aptitude test was dependent on the instructional situation.

Kaulfers continued to think about prognosis. In a paper published in 1939, he again expressed a fundamentally pessimistic view, and concluded that 'prognosis as a panacean solution to foreign-language problems is destined long to remain in the limbo of wishful thinking.' The fundamental problem as he saw it was the proliferation of approaches to teaching: 'it is inconceivable that any one test, however comprehensive, could predict achievement in a field in which such a variety of methods, materials, and objectives abound.'

In the same year, Kaulfers wrote reviews of the Symond's *Foreign Language Prognosis Test* (2:1340)[4] and the *Luria-Orleans Modern Language Test.* (2:1341) The

[4]References are to Buros (1975).

former he considered to be no more than 'a linguistically weighted intelligence test,' to lack any validity data, and to achieve too low a prediction correlation to warrant its use to reject a student. In any case, its usefulness would be limited to grammar-translation courses, and it would be too difficult for any student below eighth grade level. The second test also appeared designed to predict achievement in 'the traditional grammar-translation type of course of a decade or more ago.' He had found its validity to be low, not enough to have any advantage over more easily available measures like a twelve-minute test of English vocabulary.

Kaulfers had put his finger on a key issue: a prognosis test measured not so much a general (or even a general special) ability as a number of abilities that would be of benefit in various language learning situations. Insofar as a foreign language teaching approach was focused on the same skills that were being used in other subjects, a simple native language vocabulary test would be as good as anything else as a predictor. Aptitude, then, while a matter concerning the individual pupil, could only be defined in the context of the teaching method that was to be used.

Other pre-war studies of prognosis

The study of language aptitude and of the possibility of predicting achievement in language learning continued to be a matter of considerable academic and professional interest for the decade after the publication of Henmon *et al.* (1929).[5] It was a popular topic for theses and articles, but there was no breakthrough. Many possible predictors were investigated, such as age, attitude and personality.

In Britain, there were some beginnings of interest in prognosis in Scottish Council for Research in Education Examination Inquiry (1934) that showed that, in French, university class marks were slightly better predictors (0.69) of degree marks than were

[5]The second volume of the *Analytical Bibliography* listed seventeen items dealing with prognosis, including Walter Kaulfer's doctoral dissertation discussed above, and the third volume, covering the years 1937-1942 but its publication delayed until after the war, (Coleman, King et al., 1949) listed twenty-five items.

secondary school teachers' estimates of the Leaving Certificate Examination administered by the school (0.55).

One paper that appeared in 1939 looked ahead to much of the work that was to come. Spoerl (1939) asked what in fact constituted language learning ability? Was it intelligence, or courage, or form-color preference, or memory? She tested thirty-eight advanced German students at the American International College in Springfield, Mass., on the Henmon-Nelson test of mental ability, the Allport Ascendance-Submission Reaction Study (to test attitude and openness to suggestions in the new foreign language situation), and the Revised Minnesota Paper Form Board Test (to see if form recognition was relevant), and had them also given the Co-operative German Test. Major differences emerged between men and women: the correlation between class grade and Cooperative test score was 0.35 for men and 0.73 for women; similarly, the correlation between the intelligence measure and the grade was 0.63 for women and 0.123 for men. Neither the test of forms nor the ascendance submission test had significance relation to the German scores. Her conclusion was that while intelligence was significant for women, it was not for men.

Looking back over the first decade's work in prognosis testing, evidently the earlier expectation of Henmon and the Modern Foreign Language Study had not been met. An article by Tallent (1938) was recorded by Kaulfers as the sixtieth article published since 1901 showing that 'prognostic testing cannot be depended upon to solve foreign language problems.' Prophesy, it seemed, was dead.

A more dispassionate reconsideration suggests that the researchers of the period had helped clarify the issue enormously, and recognized the limitations of their task in that they were being asked to predict a more or less immeasurable attainment in uncontrolled and variegated learning situations. They were aware of the problems caused by the variation in goals and methods of teaching contexts, cognizant of the need for multiple rather than single predictors, and open to the complexity resulting from the fact that aptitude (however measured) was only one of a number of factors accounting for

achievement. The results of tests that they had developed, which were either slightly modified intelligence tests or mini-lessons in language, when used together with other available data, did permit a wise high school counselor to give useful advice to students identified as unlikely to succeed in formal language learning classes, and did permit responsible schools to make special provision for pupils who would be unlikely to benefit from such classes. Their tests were, as Carroll (1960) concluded when he started his own major work, 'reasonably effective in predicting success' in classes whose main objective was teaching the ability to read and translate a foreign language. They were to prove much less effective in predicting performance on more communicatively oriented programs, a challenge that was to be met by Carroll and others a quarter of a century later. But given the limited support for the research they had tackled, the high level of understanding reached during this first period deserves better recognition.

The Army UCLA aptitude study

The issue of prognosis did not die. During the war, admission to intensive language training courses in the military forces was based mainly on previous education. Frith (1953) reported at the 1953 Georgetown Round Table that the Air Force used scores on general intelligence and technical aptitude tests, possession of a high school diploma and a desire to study the language as the criteria for starting the study of Mandarin Chinese.

With the peace-time need for more economically sound approaches, the issue of which people to train became significant. Frith (1953) described trial courses conducted as screening devices at the Air Force Institute of Technology. Morgan (1953) reported that another government agency used the same approach, but Morgan himself believed and claimed to have demonstrated that an hour's careful study by a clinical psychologist of material collected with a battery of tests, including a projective "written interview questionnaire" and a personality inventory, would produce equally valid predictions.

As language training developed in the post-war years at the Army Language Training School in the Presidio of Monterey, the possibility of saving wasted time and effort persuaded the Army to fund the construction and validation of foreign language aptitude tests. The contract for the study went to three psychologists at the University of California, Los Angeles. The project, led by Roy M. Dorcus assisted by George E. Mount and Margaret H. Jones (1953) lasted from June 1950 to May 1953 and dealt with six languages, Russian, Hungarian, Serbo-Croatian, Arabic, Japanese and Mandarin Chinese.

A preliminary search of the literature produced 'no studies of value in the design of language aptitude tests for the selection of language trainees,' apart from some results of the language portions of the West Point Qualifying Examination. The report did not discuss any of the large body of pre-war work on foreign language prognosis described earlier in this chapter and it is not clear whether the authors knew of its existence and considered it irrelevant, or whether as psychologists coming to the field from outside they were unaware of foreign language testing literature that could have given them a jump start in their work. Analysis of data routinely collected at the Army Language School revealed that only pitch correlated significantly with any of the language proficiency scores, and that only for the first written and the first course oral examinations.

Nonetheless, encouraged by the high correlation between early and late language scores to believe that there must be measurable aptitude facts that could help predict later results, the team developed a list of ten 'major aptitude skills' which could be measured with a group pencil-and-paper test; this latter limitation prevented the testing of oral manipulation skills. The items chosen show a psychologist's rather than a linguist's view of the process of language learning. Perhaps if Harvard had been closer to Monterey, a more qualified research team might have been selected -- it was on the grounds of distance that John Carroll's bid for the contract was turned down. (Carroll, personal communication, 19 October 1993)

The test battery, different for each language, was administered to 150 incoming trainees in 1950 and scored at the University of California, Los Angeles, and compared

with proficiency scores on a complete battery of language proficiency tests also constructed by Army Language school staff for the study. The results of the study were disappointing. The West Point Qualifying examination continued to be the best predictor of the outcome of training, about 5-10 per cent above chance. Adding the selection tests did not improve the predictive power much. While there continued to be evidence of aptitude in the high correlation of early and late scores, the various aspects measured appeared 'to include a relatively small part of the aptitude and skill required in the learning of a language.' While still convinced of the existence of language aptitude, the researchers had failed to find a way to measure it.

This was surely not the first, not will it be the last time that experts from a related field have failed because of their lack of understanding of language and their unwillingness to start from the current state of knowledge in the field of language learning. Unhampered by knowledge of earlier work, they were able to repeat mistakes and look in the wrong places.

The prediction of success in intensive foreign language training

A much more systematic attack on the problem of language aptitude was made by John Carroll, in some years of research funded by the Carnegie Foundation and conducted at the Laboratory for Research in Instruction, Graduate School of Education, Harvard University. Carroll reiterated the economic basis for the concern, because of the expense of the intensive language programs that required eight to twelve months of full-time study and which were being offered in programs like the Army Language School at the Presidio of Monterey. An accurate measurement of foreign language learning aptitude should be able to provide a valuable screening device for costly governmental programs and minimize training failures, which ran as high as 80 per cent in one Japanese program that had been studied by Williams and Leavitt (1947).

Carroll premised his investigation on two 'propositions.' The first was that the facility to learn to speak a foreign language is 'a fairly specialized talent (or group of

talents)' independent of the traits included under 'intelligence.' The second was that it is rare enough in the general population to make it worthwhile to be selective in choosing people for expensive intensive programs. Intelligence tests, he pointed out, had been relatively unsuccessful in screening people for language training. Even with groups carefully selected for general intelligence, Frith (1953) had found that trial courses led to the rejection of as many as 75 per cent. of the students The prognosis tests tried in the 1920s and 1930s had generally been limited, Carroll noted, to pencil-and-paper testing of English language ability or work-sampling of short lessons in cognitive, intellectual aspects of formal language learning. These tests, which generally correlated quite highly with intelligence tests, were often reasonable predictors of learning to read and translate but they had less relevance to learning to speak a language in an intensive course. Dorcus and colleagues, Carroll graciously suggested, had 'just missed' measuring the crucial abilities, in that their tests failed to tap the relevant abilities. Memory for digits, for instance, which they tested, was not relevant to language learning, while memory for sound, which they did not test, probably was significant.

Carroll started with an initial battery that contained twenty separate tests, each intended to check one of five factors of verbal ability that had been proposed by French (1951): verbal knowledge, word fluency (knowledge of orthographic habits), fluency of expression, associative memory, and naming. Also included was a Phonetic Discrimination task developed by Stanley Sapon that asked the subject to identify the odd sound out in a triad.

Carroll tried several kinds of work-sample tests. One was an artificial language test in which subjects learned the names of a simple foreign language number system. Another was a tape recording with accompanying film strip that taught a simple artificial language. A third presented a more formal artificial language through grammar lessons.

In this approach, Carroll was working on the same double strategy followed by earlier aptitude testers. If he could, he wanted to find tests that tapped the most basic

abilities in language learning, the discrete primary skills. Failing this, he sought to find the smallest trial learning situation that would predict performance in a full course.

The new tests were tried in a number of situations. In February 1954, 111 men pre-screened for admission to an eight month intensive course offered for the U.S. Air Force at Yale University took a four hour battery of tests. They then went into a three day preliminary training period, during or after which thirty-one withdrew voluntarily. The validity analysis was based on the remaining eighty, only thirty-three of whom were selected for the full course. Using as the criterion measure either grades given by instructors or the selection decision, a large number of test variables showed significant correlations. The summed results of four tests (artificial language learning, phonetic association, words in sentences, and paired associates) produced a multiple R of 0.74. The prediction test and the trial course had agreed in sixty-six out of eighty cases.

A second trial was carried out in June 1954, using some new types of items. Once again, validity coefficients were remarkably high, a multiple R of 0.77 -- and, using some of the new tests, 0.839. On the basis of these successes, the *Psi-Lambda*[6] *Foreign Language Aptitude Battery* was made available to the Air Force in 1955 for further testing, with generally satisfactory results. The screening policy finally adopted by the Air Force was to use the result of the battery as a criterion for admission to the trial course, and make a further cut after that.

Two series of tests were conducted to check the relevance of the battery for different types of languages. While the correlation in one sample was lowest in predicting success in learning languages with characters (Japanese, Chinese, Korean), this did not show up in a second sample. This result and other analyses supported the hypothesis of the non-specificity of language aptitude. The battery seemed to predict oral and written skills equally well, depending on the instructional approach.

---

[6]An abbreviation, Carroll noted, for psycholinguistic.

Experimental testing was also conducted at the Foreign Service Institute of the U.S. Department of State. Good correlations (about 0.70) were found with instructor grades in six-month long courses in twelve different languages. In another test, eighty-three trainees at the Foreign Service Institute were given the battery, which achieved a multiple R of 0.778 with performance at the end of a six month course. The test was much better than the prediction based on a fifteen-minute 'diagnostic interview' given to the candidates by the chair of the language department in which he was to study. The results of this study also produced evidence of the effect of age; while the subjects' aged showed a slightly negative linear correlation with their success in language learning, the fact that adding the age variable to the aptitude test did not improve the prediction showed that the aptitude test measured whatever in the age variable was relevant to success in language learning; it further contradicted the notion that older people cannot learn foreign languages successfully.

Carroll (1960) reported two situations in which the aptitude battery failed to make significant predictions. Sixty two persons in six month courses conducted by the National Security Agency were given a battery of tests before they began courses (typically six months long); the tests failed to predict their grades in these courses, which were concerned with the use of foreign language skills in "cryptanalysis and related matters." Carroll explained this as a result of the criterion being "poorly defined" or "irrelevant."(It is likely that Carroll was given no further details of the course or of the criterion tests. The National Security Agency tended to be security-conscious; as I recall, its linguists used to pretend to be working for the CIA.) In the second case of failure that he reported, the battery was given to two classes of U.S. Air Force personnel learning Russian in an intensive program in a charitably unnamed American university. Carroll attributed the lack of correlation between the battery and the criterion grades to the inconsistency of the latter scores, as well as to such associated matters as "the quality of the teaching, the quality of the text materials, and the reliability of the grading." From all these studies,

Carroll was satisfied that he had good evidence that the tests in the battery were "generally speaking, highly valid."

The Modern Language Aptitude Test

Given the general success of the battery, a commercial form of the Carroll and Sapon test was published in 1959 by the Psychological Corporation under the name, *Modern Language Aptitude Test.* In this form, it was tried out in the summers of 1958 and 1959 with students in intensive eight week summer courses in Arabic, Persian, Turkish or Modern Hebrew, producing correlations of about 0.5 with final grades.

In a major paper reviewing his work in developing successful aptitude measures, Carroll (1960) raised a more fundamental question. His studies to date had assumed that success was a direct function of measured aptitude. Such a model was 'oversimplified, if not downright wrong.' A better model would take into account other relevant factors, such as motivation and instructional variables. He proposed a model that included at least two instructional variables (adequacy of presentation and the time allowed for learning) and three individual variables (verbal intelligence, aptitude -- or amount of time needed to learn -- and motivation -- or the amount of time the learner would apply himself to the task. Using the resulting model, Carroll was able to demonstrate how variation in the conditions of the various courses accounted for variation in the predictive ability of the aptitude battery. Because aptitude is not the only variable accounting for success in language learning, its validity can only be shown when the other factors are taken into account.

In summing up his major study, Carroll concluded that language aptitude consisted of the four distinct and measurable abilities: phonetic coding[7] -- the ability to code an auditory phonetic signal so that it could be remembered for more than a few seconds, grammar handling[8]-- the ability to recognize functions of words in sentences, rote

[7]The Phonetic Coding Factor, Carroll (1993:171) notes, may be identical to the Spelling Cluster of abilities.

[8]It is still not clear, Carroll (1993:176) remarks, if the Grammatical Sensitivity factors represent a learned ability.

memorization ability of a large number of foreign language items,[9] and inductive language learning ability.[10] With the completion of this major body of research, then, Carroll could feel reasonably confident that he had managed to identify and measure the chief factors involved in aptitude for learning to speak a foreign language. His tests were able to account for most of the variation that could reasonably be attributed to aptitude.

While Carroll and Sapon's work did include validation of the use of the test in high school situations, the main goal of their test was to predict success in intensive courses of the kind more likely to be used at university level or for adults. A number of years later, Paul Pimsleur translated his findings into a published test battery, The *Pimsleur Language Aptitude Battery*.

The state of prophecy

When the Temple was destroyed, the Talmud says, the power to predict the future was taken away from prophets and given to fools and children.[11] Henmon and his colleagues' initial hope of achieving close to perfect prognosis was, it is now clear, over-optimistic. But they managed to show, and Pimsleur confirmed, that verbal intelligence tests do a good job in predicting not just how well a student will do at school, but how well he or she will do in typical foreign language classes, making it possible to schools to exclude students who are probably going to fail.

John Carroll added three vitally important dimensions. First, more successfully than anyone, he developed tests that measured, as well as anything can, some of the components of individual variation in ability to learn to speak a foreign language. The items in the *Modern Language Aptitude Test* continue to show up as robust factors in

---

[9]The memory factors identified in the aptitude studies appear to be special. See Carroll (1993:297-298).

[10]A more general foreign language ability factor may emerge, Carroll (1993:176-7) now says, if the test battery does not permit the Grammatical Sensitivity and the Phonetic Coding factors to emerge.

[11]Babylonian Talmud, Tractate *Baba Bathra*, 12b

studies of second language learning.[12] Second, he proposed a model that showed how measurable abilities interact with goals and methods. Third, his extended model made the whole issue clearer, by showing that aptitude was only one of the factors involved in what I have called a general theory of second language learning (Spolsky 1989).

Ultimately, then, the work on prognosis in the 1920s and 1930s and on language aptitude in the 1950s produced tests that could be used cautiously for selecting promising language students, and it provided, perhaps more important, an improved understanding of the nature of second language learning. Aptitude, this work clearly showed, is only one of the factors that can be used to predict success in second language learning. In seeking to make further advances in the field, it is unwise not to build on the work of our predecessors.

References

Buros, O. K. (ed.) 1975. *Foreign language tests and reviews*. Highland Park, New Jersey, The Gryphon Press.

Carroll, J. B. 1960. The prediction of success in intensive foreign language training (final revision). Laboratory for Research in Instruction, Graduate School of Education, Harvard University.

Carroll, J. B. 1962. 'The prediction of success in intensive foreign language training' in R. Glaser (ed.): *Training research and education*. Pittsburgh, The University of Pittsburgh Press. 87-136.

Cheydleur, F. D. 1932a. 'Mortality of modern languages students: its causes and prevention.' *Modern Language Journal* 17(2): 104-136.

[12]For example, in a study of language gains by 658 American students in four month study-abroad programs in Russia, Ginsberg (1992) found two MLAT tests show up as significant predictors for gains in listening and reading. Carroll (1993) provides a reanalysis of early studies. The *Modern Language Aptitude Test* is still, at this writing, in print and use.

Coleman, A. and C. B. King (ed.) 1938. *An analytical bibliography of modern language teaching, vol. II, 1932-1937*. Chicago, University of Chicago Press.

Dorcus, R. M., G. E. Mount, and M. H. Jones. 1953 (mistakenly dated 1952). Construction and validation of foreign language aptitude tests. University of California, Los Angeles, for the Adjutant General's Office.

Edgeworth, F. Y. 1888. 'The statistics of examinations.' *Journal of the Royal Statistical Society* **51**: 599-635.

Foucault, M. 1975. *Surveiller et punir: naissance de la prison*. Paris, Gallimard.

Foucault, M. 1979. *Discipline and punish: the birth of the prison*. New York, Vintage.

French, J. W. 1951. *The description of aptitude and achievement tests in terms of rotated factors*. Chicago, University of Chicago Press.

Frith, J. R. 1953. 'Selection for language training by a trial course' in A. A. Hill (ed.): *Report of the fourth annual roundtable meeting on languages and linguistics*. Washington, DC, Institute of Languages and Linguistics, Georgetown University. 10-15.

Henmon, V. A. C. 1929. *Achievement tests in the modern foreign languages, prepared for the Modern foreign language study and the Canadian committee on modern languages*. New York, The MacMillan company.

Henmon, V. A. C., J. E. Bohan, C.C. Brigham, L.T. Hopkins, G.A. Rice, P.M. Symonds, J.W. Todd, and R.J. Van Tassel (ed.) 1929. *Prognosis tests in the modern foreign languages: Reports prepared for the Modern Foreign Language Study and the Canadian Committee on Modern Languages*. Publications of the American and Canadian Committees on Modern Languages. New York, The MacMillan Company.

Hunt, T., F. C. Wallace, S. Doran, K. C. Buynitzky, and R. E. Scharz. 1929. *Language Aptitude Test: George Washington University*. Washington, DC, Center for Psychological Service, George Washington University.

Kaulfers, W. V. 1931. 'Present state of prognosis in foreign languages.' *School and Society* **39**(8): 585-596.

Kaulfers, W. V. 1933a. Forecasting efficiency of current bases for prognosis. Unpublished doctor's dissertation, Stanford University.

Kaulfers, W. V. 1939. 'Prognosis and its alternatives in relation to the guidance of students.' *German Quarterly* **12**(3): 81-84.

Latham, H. 1877. *On the action of examinations considered as a means of selection*. Cambridge, Deighton, Bell and Company.

Lau, L. M. 1933. The use of the Symonds' Foreign Language Tests in Beginning French. Unpublished master's thesis. University of Chicago.

Luria, M. A. and J. S. Orleans 1928. *Luria-Orleans Modern Language Prognosis Test.* Yonkers, N. Y., World Book Company.

Michel, S. V. 1934. Prognosis in the modern foreign languages. Unpublished master's thesis, University of Minnesota.

Michel, S. V. 1936. 'Prognosis in German.' *Modern Language Journal* **20**(5): 275-287.

Morgan, W. J. 1953. 'A clinical approach to foreign language achievement' in A. A. Hill (ed.): *Report of the fourth annual roundtable meeting on languages and linguistics*. Washington, DC, Institute of Languages and Linguistics, Georgetown University. 15-21.

Richardson, H. D. 1933. 'Discovering aptitude for the foreign languages.' *Modern Language Journal* **18**(3): 160-170.

Scottish Council for Research in Education Examination Inquiry 1934. *The prognostic value of university entrance examinations in Scotland*. London, University of London Press, Ltd.

Spoerl, D. T. 1939. 'A study of some of the possible factors involved in foreign language learning.' *Modern Language Journal* **23**: 428-431.

Spolsky, B. 1989a. *Conditions for second language learning: introduction to a general theory*. Oxford, Oxford University Press.

Spolsky, B. 1995. *Measured Words*. Oxford, Oxford University Press.

Stoddard, G. D. and G. E. Vander Beke 1925. *Iowa Placement Examinations: Foreign Language Aptitude*. Iowa City, State University of Iowa.

Symonds, P. M. 1930a. Foreign Language prognosis test. New York, Teachers College, Columbia University.

Symonds, P. M. 1930b. 'A foreign language prognosis test.' *Teachers College Record* **31**: 540-546.

Tallent, E. R. E. 1938. 'Three coefficients of correlation that concern modern foreign languages.' *Modern Language Journal* **22**(8): 591-594.

Williams, S. B. and H. J. Leavitt 1947. 'Prediction of success in learning Japanese.' *Journal of Applied Psychology* **31**: 164-168.